ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

Predictive Driver Monitoring Using Multimodal AI for Road Safety

Arvind Kumar[1]; Arpan Mukherjee[2]

[1]Industrial Autonomy & Engineering Unit, Tata Consultancy Services Ltd, Noida, India.
[2]Department of Chemical Engineering, Indian Institute of Technology Madras Chennai, India

[1]E-mail: arvind.jki@gmail.com

Publication Date: 2026/03/09

Abstract: Driver behavior remains a leading factor in road accidents, yet existing monitoring systems typically rely on single data modalities such as facial expressions or speech alone, limiting their reliability and contextual awareness. This work proposes a comprehensive driver behavior monitoring system using multimodal AI, uniquely integrating video, audio, and vehicle speed telemetry — an approach that remains underexplored in existing literature — to predict driver emotions and behaviors in real time. The system analyzes facial cues to detect visual anomalies, processes audio inputs to infer emotional states, and incorporates speed telemetry to provide additional behavioral context. This fusion of modalities is designed to improve classification accuracy and reduce false positives compared to unimodal approaches. Performance evaluation is conducted using benchmark datasets for both video-based and audio-based emotion recognition, with comparative analysis between individual and combined modalities. By addressing the challenges of multimodal integration and real-time processing, this research contributes a novel and effective framework for intelligent driver assistance systems, advancing the goal of enhanced road safety through predictive behavioral intervention. Additionally, this research is being extended to incorporate an intermediate-fusion-based multimodal decision framework, wherein top predictions from image, audio, and vehicle telemetry are jointly fed into a decision model (e.g., Random Forest or SVM) for improved context-aware warning generation. This approach addresses prior limitations of ensemble-style fusion and better aligns with the goals of true multimodal AI.

Keywords: *Multimodal Driver Monitoring; Real-Time Emotion Detection; Deep Learning; Audio-Visual Data Fusion; Road Safety; Predictive Behavior Analysis.*

How to Cite: Arvind Kumar; Arpan Mukherjee (2026) Predictive Driver Monitoring Using Multimodal AI for Road Safety. *International Journal of Innovative Science and Research Technology*, 11(3), 87-96. https://doi.org/10.38124/ijisrt/26mar176

I. INTRODUCTION

Road safety is a critical global concern, with driver behavior identified as a leading contributor to traffic accidents, especially among young adults [1, 2]. Traditional Driver Monitoring Systems (DMS) often rely on single data modalities such as video or audio, which limits their ability to capture the complexity of driver states. For instance, video-based systems primarily analyze facial cues [3], while audio systems interpret vocal signals to infer emotional tone [4]. However, these unimodal approaches struggle to account for the interplay between visual, auditory, and behavioral signals that influence driving safety. To address this, recent research has turned toward Multimodal AI, which integrates inputs from video, audio, and even vehicle telemetry to enhance contextual understanding [5]. This integration has been shown to improve classification accuracy and reduce false positives, enabling more reliable real-time interventions.

Despite these advancements, existing DMS face several challenges. These include their dependency on unimodal data, a lack of synchronized multimodal datasets, difficulties in real-time processing of high-volume data, and limited adaptability across diverse driving environments [6]. While individual models for facial and speech emotion recognition have matured, gaps remain in the real-time fusion of multimodal data and the availability of datasets that reflect real-world variability [7, 8].

This research addresses the need for a robust multimodal driver monitoring system that integrates video and audio streams to predict unsafe behaviors in real time. A distinctive feature of this work is the use of vehicle speed data—extracted directly from video frames—as an additional contextual signal, which, to the best of our knowledge, has not been adequately explored in prior literature.

The proposed system processes video and audio data in parallel. Video analysis detects visual anomalies and extracts vehicle speed, while audio processing infers emotional states such as anger, fear, and distraction. These outputs are then combined to assess driver behavior comprehensively. This

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

architecture is designed to enable real-time intervention, ensure computational efficiency, and address key challenges in multimodal data fusion. The primary objectives of this work are to develop a multimodal driver monitoring framework, evaluate its performance on benchmark datasets, and assess its advantages over unimodal systems for practical deployment in intelligent transportation systems. While this paper primarily describes an ensemble-based decision protocol combining outputs from image and audio modalities, work is extended into a more integrated multimodal framework. Specifically, feature-level fusion of top predictions from visual and acoustic models—along with vehicle telemetry data—is being explored using a meta-classifier (Random Forest). This aims to capture inter-modality dependencies and enhance decision robustness.

II. RELATED WORK

Driver behavior monitoring systems play a crucial role in improving road safety by identifying risky behaviors such as distraction, fatigue, and emotional instability. These behaviors are recognized as major contributors to road accidents by the World Health Organization (WHO) and Highway Traffic Safety Administration (NHTSA). To address this, recent research has focused on developing intelligent monitoring systems that use facial and vocal inputs to detect the driver's emotional and cognitive states in real-time.

Facial Emotion Recognition (FER) is one of the primary techniques used in such systems. Convolutional Neural Networks (CNNs) have proven effective in capturing spatial features from facial images, making them a preferred choice for static image-based emotion detection [3, 5]. However, CNNs often face challenges with real-time performance and are sensitive to occlusions and lighting variations. To address these limitations, *Squeeze-and-Excitation Networks* (*SENet)* have been introduced [9]. SENets enhance CNN performance by recalibrating channel-wise features, improving sensitivity to fine-grained emotional cues [6]. Nonetheless, their increased computational cost can hinder deployment on low-power embedded systems. Hybrid models such as CNN-LSTM further improve recognition accuracy by incorporating temporal information from video sequences. These models are particularly suited for dynamic emotion recognition but demand large datasets and greater training time, which complicates real-time application [4, 10].

Speech Emotion Recognition (SER) complements facial analysis by providing clues about stress, fatigue, or agitation through vocal characteristics. Traditional SER systems utilize Mel-Frequency Cepstral Coefficients (MFCCs) to extract features relevant to human auditory perception [11]. These features are computationally efficient but may degrade in noisy conditions. Machine learning models like Radial Basis Function (RBF) networks have shown promise due to their fast training and simplicity, although they struggle with generalization on larger and more diverse datasets [12]. Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks have gained popularity for handling sequential audio data, offering better performance in capturing temporal dependencies [7]. GRUs, in particular, are a lighter alternative to LSTMs, making them suitable for systems requiring lower computational overhead.

Several studies have explored the benefits of integrating both FER and SER into a unified framework to create robust multi-modal driver monitoring systems. This integration enhances the accuracy and reliability of emotion recognition, as it draws from both visual and auditory inputs [2]. However, such systems face challenges including synchronization issues, higher data processing requirements, and increased system complexity.

In summary, a variety of approaches have been proposed to enhance driver emotion detection. CNNs and SENets are well-suited for spatial feature extraction from facial data, while GRUs and MFCC-based models excel in sequential and auditory analysis. Although multi-modal systems yield higher accuracy, they introduce trade-offs in terms of computational demands and real-time feasibility. Therefore, selecting the most appropriate method depends on the specific constraints and goals of the application.

III. METHODOLOGY

The proposed system aims to develop a predictive driver behavior monitoring and real-time intervention framework utilizing multimodal AI. By integrating both video and audio data, the system is designed to detect unsafe driver behaviors and emotional states, thereby enabling timely interventions to enhance overall road safety. This section specifically outlines the methodology employed for developing and training image-based and audio-based models for predictive driver behavior monitoring, with a primary focus on facial emotion recognition.

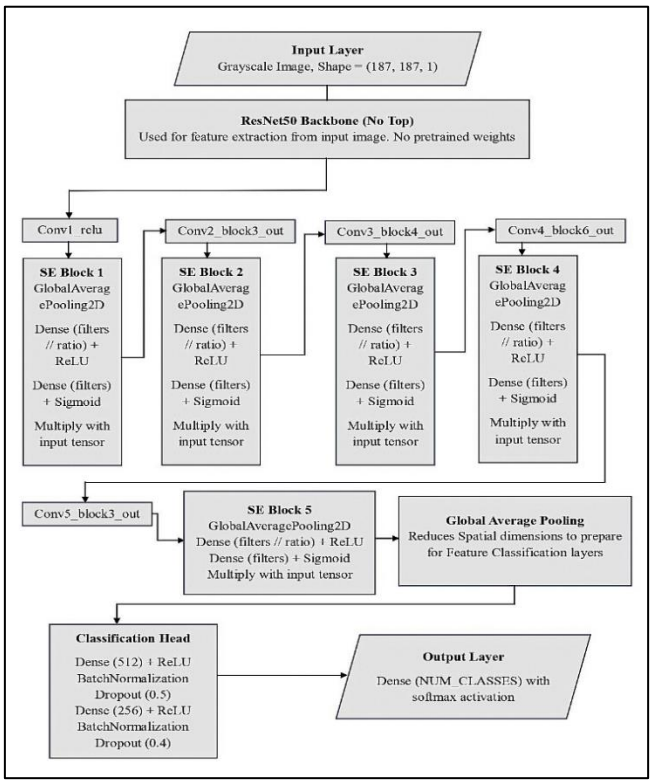

Fig 1 Simplified Se-ResNet50 Model

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

For the image-based analysis, a *Se-resNet50* architecture is selected. The simplified architectural flow of the Se-ResNet50 model is presented in Fig. 1. This model processes driver facial expressions captured from in-cabin video footage and classifies them into one of seven categories: *Drowsy*, *Visually Distracted*, *Cognitively Distracted*, *Angry*, *Fidgety*, *Fearful*, and *Neutral*. Before training, input data underwent several preprocessing steps. All image frames were standardized to *JPEG* format and converted to *grayscale*, reducing the computational burden while retaining essential features for emotion detection. Each frame was resized to 187x187 pixels, ensuring consistent dimensionality. Pixel intensity normalization was applied by scaling values between 0 and 1 to facilitate stable and efficient training. To augment data and improve generalization, random transformations such as rotation, flipping, and zooming were applied.

The Se-ResNet50 model architecture is a hybrid deep learning model that combines the strengths of ResNet50 and Squeeze-and-Excitation (SE) blocks to improve feature learning and classification performance (Fig. 1). The model leverages the residual learning capabilities of ResNet50 to capture deep hierarchical features while embedding SE blocks at multiple stages throughout the network. These SE blocks enhance the network's sensitivity to essential features by adaptively recalibrating channel-wise feature responses. This recalibration process involves learning how much importance to assign to each channel, enabling the model to emphasize more relevant features and suppress less useful ones.

After extracting features through a series of convolutional and recalibrated layers, the architecture transitions into a global pooling operation that compresses the spatial dimensions. This is followed by a fully connected structure designed for classification: two dense layers with progressively reduced units, each followed by batch normalization and dropout for regularization and better generalization. The final classification is performed using a SoftMax layer, which outputs the probability distribution across the target classes. This architectural design makes the model well-suited for tasks with nuanced patterns or class imbalances, as it combines strong feature extraction with focused attention mechanisms.

In parallel, the audio processing module focuses on inferring emotional states from vocal input, which provides complementary information to facial cues. The audio pipeline processes raw speech data recorded during driving sessions. Audio features are extracted using *Librosa*'s *melspectrogram* function with 128 mel bands, followed by logarithmic scaling to enhance perceptual relevance and normalization to ensure input consistency. Each resulting spectrogram is either padded or truncated to 200 time frames to maintain uniform sequence length across the dataset.

The architecture selected for the audio model is a hybrid structure combining convolutional layers with a *Bidirectional Gated Recurrent Unit (BiGRU)* and an *attention mechanism*. The architectural flow of the hybrid CNN–SE–BiGRU attention model is presented in Fig. 2. This configuration is chosen based on its superior performance in capturing both local time-frequency features and long-range temporal dependencies. The input to the model is a 2D tensor of shape (200,128,1), representing time, frequency, and channel dimensions. Three convolutional blocks are used to extract features, with 32, 64, and 128 filters, respectively, and 3x3 kernel sizes. Each convolutional block includes *ReLU* activation, *BatchNormalization*, and *MaxPooling* layers. *Squeeze-and-Excitation (SE)* blocks are added after the second and third convolutional stages to recalibrate channel-wise feature maps, allowing the network to emphasize more informative features.

Following the convolutional layers, the spectrogram is reshaped into a 2D sequence format and passed into a *Bidirectional GRU* layer, which processes the temporal dynamics from both forward and backward directions using 64 units in each direction. To enhance interpretability and focus the network on the most informative parts of the sequence, an *attention mechanism* is applied to generate a context vector that weighs the importance of each time step. This context vector is passed through a fully connected dense layer with 128 units and *ReLU* activation. *Dropout* with a rate of 0.5 is applied to mitigate overfitting. The final output layer is a *SoftMax* classifier that predicts among three emotional states: *Anger*, *Fear*, and *Neutral*.

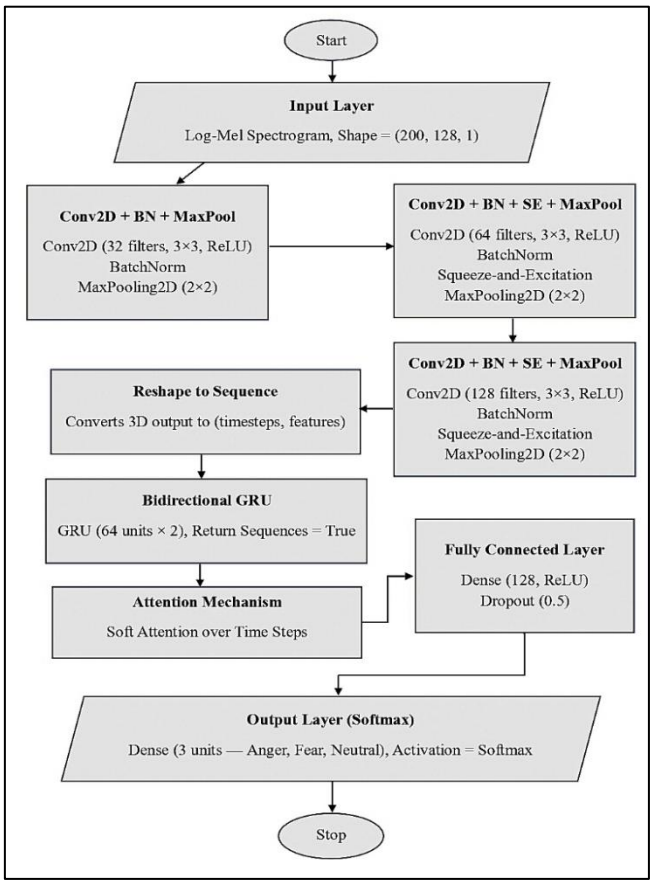

Fig 2 Hybrid CNN–SE–BiGRU Attention Model

Class imbalance is an implicit challenge associated with real-time datasets [13]. To address class imbalance in the audio dataset, class-specific data augmentation techniques are applied. For instance, *pitch shifting*, *time stretching*, and *noise*

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

injection are used for underrepresented classes like *Fear* and *Neutral*, while *SpecAugment* is applied exclusively to the *Fear* class to improve the model's robustness to intra-class variability. A custom *Focal Loss* function, with parameters $\alpha=2.0$ and $\gamma=3.0$, is used to further address imbalance by penalizing well-classified examples and emphasizing harder cases. Training is conducted using the *Adam* optimizer with an initial learning rate of 0.001. The training loop employed *EarlyStopping* and *ReduceLROnPlateau* callbacks to prevent overfitting and adjust the learning rate dynamically. The model is trained for a maximum of 120 epochs with a batch size of 32, using both original and augmented samples. The final model is exported in *.h5* format for future inference and deployment.

During real-time testing, the system performs second-wise processing of input video and audio. A representative frame is extracted per second and passed to the Se-ResNet50 model for visual analysis, while the corresponding audio is processed using the CNN-SE-BiGRU model. Starting from the second-second, audio input is handled using a sliding window of two seconds to provide temporal context. The vehicle's speed data is also logged and used for behavioral context. Model outputs are fused using a decision protocol that compares prediction probabilities. If both modalities predict the same behavior or emotion, the one with higher confidence triggers a voice warning. If predictions conflict, the video model is prioritized. A predefined priority list of behaviors resolves ties in prediction probability. This logic enables the system to issue real-time, behavior-aware interventions.

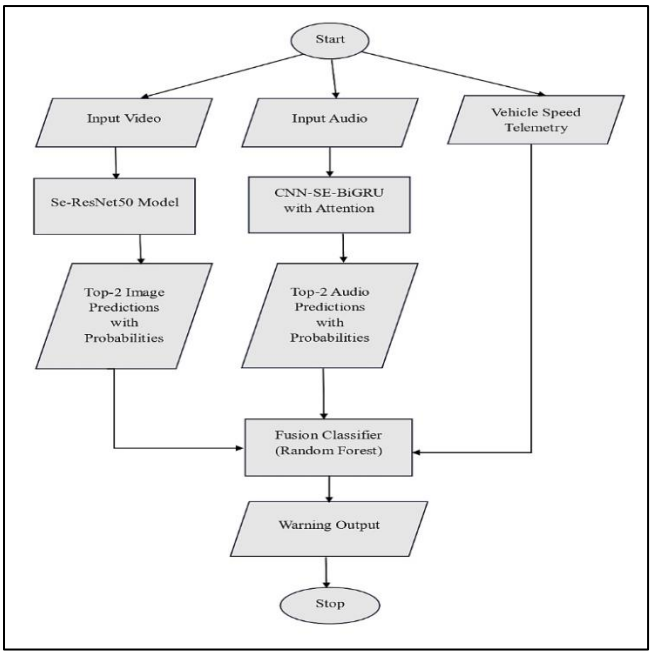

Fig 3 Intermediate Fusion Model for Warning Generation

The current system utilizes a rule-based late fusion strategy that compares outputs from individual models and makes decisions based on predefined heuristics. While this method provides a clear and interpretable decision framework and treats each modality—such as visual and audio inputs—with equal importance, it does not fully exploit the potential interdependencies between them. To enhance this, a new multimodal decision pipeline is being developed to create a more integrated and adaptive approach. In this design, a unified feature vector is constructed for each time frame, combining the top two predictions (with their probabilities) from the Se-ResNet50 image model, the top two predictions (with probabilities) from the CNN-SE-BiGRU audio model, and the vehicle's speed as a contextual numeric feature. These combined features are input into a supervised classifier—specifically, a Random Forest—which is trained to predict appropriate warnings or classify driver behavior and emotional state [14]. By learning from historical data, the classifier can adaptively weigh each modality, allowing it to make more context-sensitive and accurate decisions. This new approach maintains the clarity of the original system while introducing the ability to learn joint representations across modalities. Training will be conducted using the same test videos, for which expert-annotated warning labels are being prepared, marking a significant step toward a more robust and truly multimodal AI pipeline. The approach is illustrated in Fig. 3.

IV. EXPERIMENTAL SETUP

This research utilizes a *multimodal dataset* and a *moderately powered computational environment* to develop, train, and evaluate the proposed predictive driver behavior monitoring system. The dataset is divided into three core types: image, audio, and custom test datasets. The image dataset is compiled from publicly available sources, emphasizing facial expressions and driver behaviors. It includes contributions from *AffectNet* for facial emotion recognition [15], the Drowsy Detection Dataset for fatigue-related behavior, and the State Farm Distracted Driver Detection Dataset for distraction-related activities [16].

From *AffectNet*, images annotated with emotions such as *Anger*, *Fear*, *Happy*, *Neutral*, *Sad*, *Surprise*, *Disgust*, and *Contempt* were obtained. However, the *Contempt* class was excluded due to its limited relevance in driver behavior monitoring, and a new *Drowsy* class was integrated using images from the *Drowsy Detection Dataset*. The drowsy dataset was created by extracting frames from video sequences of drivers in various fatigue states, with facial regions identified and labeled according to drowsiness intensity.

In addition, distraction-related behaviors such as *cognitively distracted*, *visually distracted*, and *fidgetiness* were derived from the *State Farm* dataset. These new classes were formed by grouping driver activities such as phone usage, texting, reaching behind, or manipulating objects while driving. After preprocessing, which included converting images to grayscale, resizing to 187x187 pixels, and storing them in *.jpg* format, a final dataset of 11 emotion and behavior classes was prepared. Out of these, 7 classes—*Anger*, *Drowsy*, *Fear*, *Cognitively Distracted*, *Visually Distracted*, *Fidgetiness*, and *Neutral*—were selected for model training, as they align most closely with real-time driver monitoring goals aimed at enhancing road safety. A validation set was curated from the dataset to monitor performance during training, ensuring effective generalization and minimizing overfitting.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

The audio dataset focuses on emotion recognition from speech, providing complementary insight into the driver's state. Sourcing speech-based emotional recordings was challenging due to the scarcity of datasets matching driving scenarios. After extensive review, the *RAVDESS* dataset was selected as the primary source. It features audio samples from professional actors across eight emotion categories, out of which *Neutral*, *Angry*, and *Fearful* samples were chosen for their relevance to in-vehicle emotion detection. To further diversify the speech data and mitigate class imbalance, supplementary datasets including *CREMA-D*, *SAVEE*, and *TESS* were incorporated. Only audio samples corresponding to the three target emotions were used to maintain consistency with the goals of this system.

All audio clips were preprocessed using dataset-specific techniques suitable for spectrogram generation and padded or truncated to consistent lengths for uniform input to the neural network models. Speech-based validation was similarly performed to reduce loss and improve emotion classification precision. In the absence of any publicly available multimodal dataset combining driver video and audio in real-time, a *custom dataset* is developed for testing. This self-recorded dataset simulates realistic driving conditions, combining driver facial video and simultaneous speech. Frames are converted to grayscale and resized to 187x187 pixels, with any blank or overexposed frames removed for quality assurance. The audio track from these videos is separately processed using the same preprocessing methods used during training, as discussed in the model design methodology.

All model development, training, and testing tasks were carried out on a personal machine running a 64-bit *Windows 11 Home Single Language* operating system, version 24H2, with OS build 26100.4061 and *Windows Feature Experience Pack* 1000.26100.84.0. The system is powered by an *11th Generation Intel (R) Core (TM) i7-1165G7* processor clocked at 2.80 GHz, with support for *x64-based architecture*. The machine is equipped with 16 GB of installed RAM (15.8 GB usable), allowing for smooth execution of training and evaluation cycles for the moderately sized datasets used in this study. While the system does not include dedicated GPU acceleration, optimizations such as *batch processing*, *data generators*, and careful management of training cycles through callbacks like *EarlyStopping* and *ReduceLROnPlateau* were implemented to ensure resource efficiency and prevent overfitting. All models were exported in *.h5* format for seamless inference and future deployment.

For training the new fusion classifier, synthetic multimodal feature vectors are being generated from the custom video test dataset. Each second of the test video is processed to extract top image and audio predictions along with vehicle speed, and expert labels are assigned to indicate the correct behavioral warning (if any). These labeled examples will be used to train and validate the new fusion model.

V. RESULTS

The evaluation of the proposed models employs standard classification metrics: accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a comprehensive understanding of the models' performance, particularly in handling class imbalances and identifying specific driver behaviors.

The results of the *Se-ResNet50* model are presented in Table 1. The *Se-ResNet50* model demonstrates robust performance across most classes. It achieves an overall accuracy of 92%, indicating a high rate of correct classifications. Precision and recall values are notably high for classes such as *cognitively distracted*, *drowsy*, *fidgetiness*, and *visually distracted* behaviors, reflecting the model's effectiveness in recognizing these specific states. The *macro* and *weighted average F1-scores* stand at 0.91 and 0.92 respectively, underscoring the model's consistent performance across all classes, even in the presence of *class imbalances*.

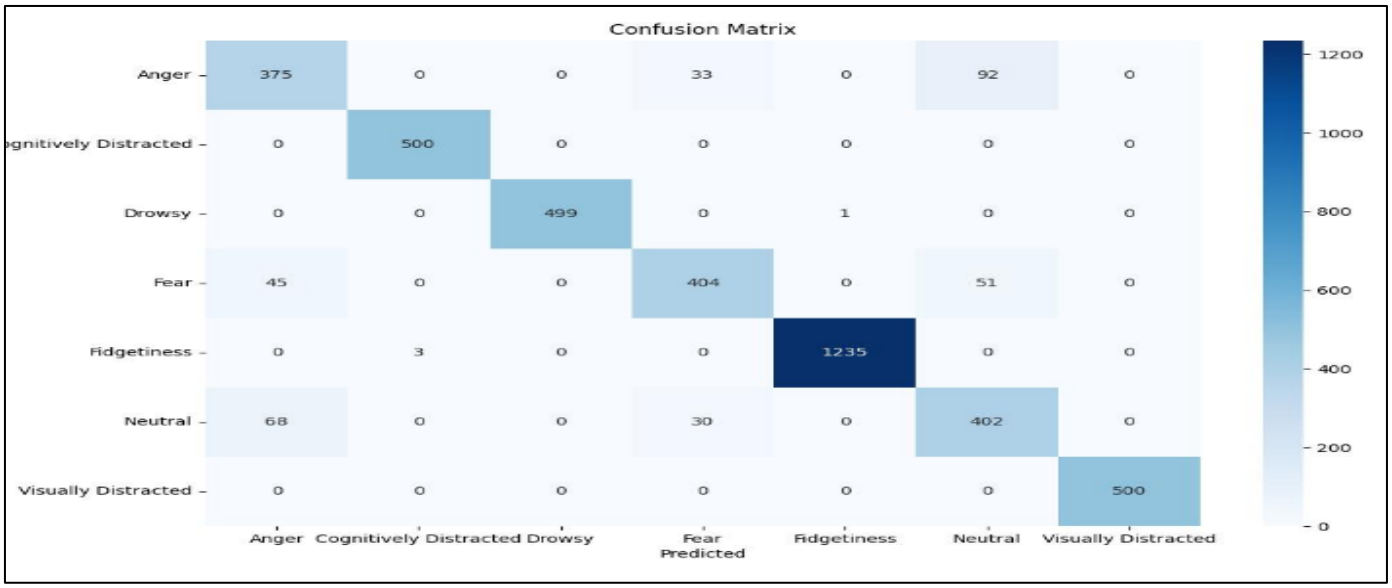

Fig 4 Confusion Matrix of the Se-ResNet50 Model

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

The confusion matrix for the Se-ResNet50 (Fig. 4) model reveals strong classification performance for most categories, with the highest accuracy in the *Fidgetiness* class, showing 1233 correct predictions and minimal misclassifications. Similarly, the model performed perfectly on *Cognitively Distracted*, *Drowsy*, and *Visually Distracted*, each with 500 correct predictions and no false positives or negatives (except one misclassification from *Visually Distracted* to *Fidgetiness*). Some misclassifications occurred between emotionally similar or visually overlapping classes—for instance, 93 *Neutral* images were misclassified as *Anger*, and 51 *Fear* images were predicted as *Anger*. This suggests the model could benefit from improved discrimination between nuanced emotional states.

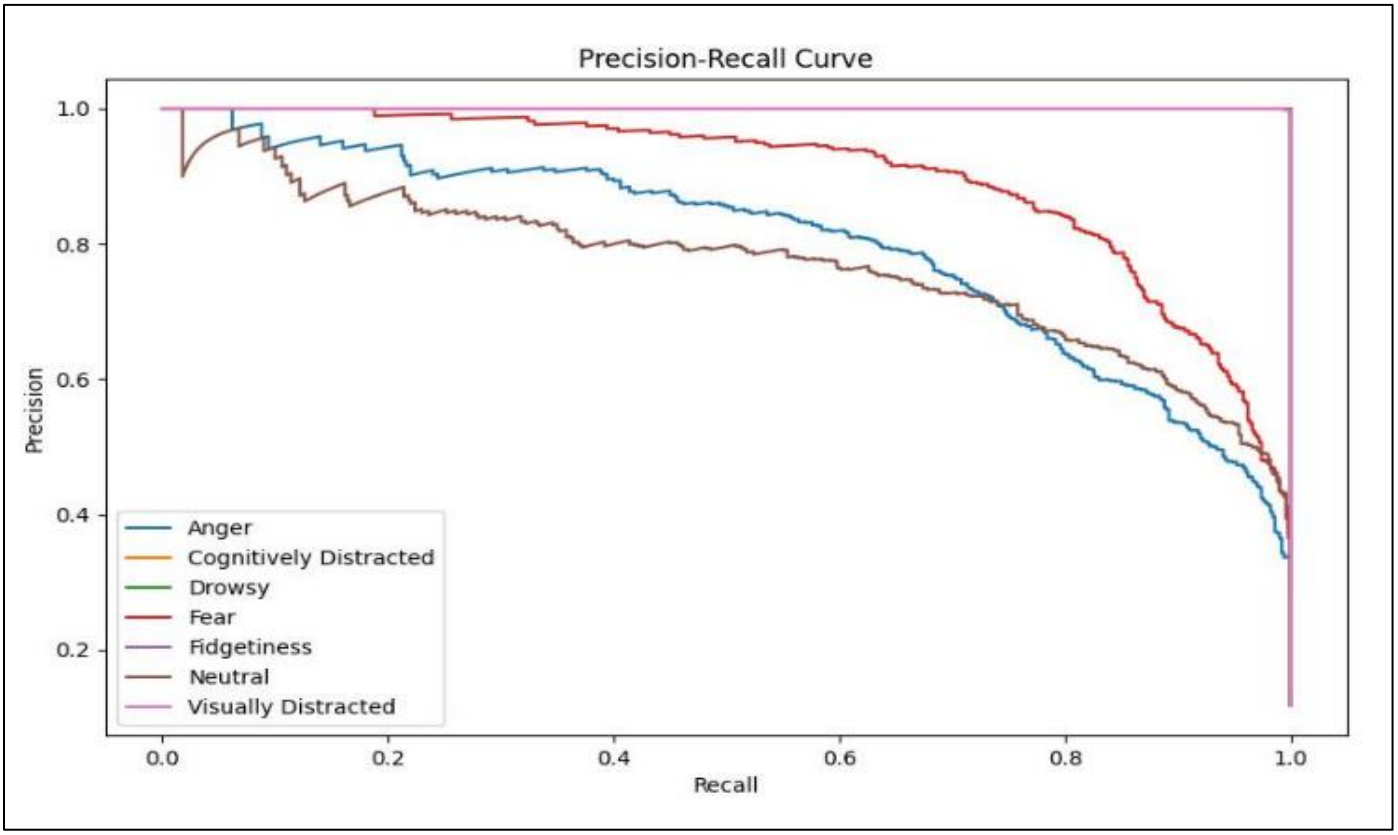

Fig 5 Precision-Recall Curves for Se-ResNet50 Model

The Precision–Recall and ROC–AUC curves for the Se-ResNet50 model are shown in Fig. 5 and Fig. 6, respectively. The ROC-AUC and Precision-Recall curves further confirm the model's strong overall performance. The ROC-AUC scores for most classes, including *Cognitively Distracted*, *Drowsy*, *Fidgetiness*, and *Visually Distracted*, reached a perfect 1.00, indicating excellent class separability. *Fear* also showed very strong performance with an AUC of 0.98, while *Anger* and *Neutral* maintained respectable AUCs of 0.97. The precision-recall curves reinforce this, especially with near-perfect precision and recall in classes like *Cognitively Distracted* and *Visually Distracted*. However, a slightly lower curve for *Anger* and *Neutral* again points to challenges in distinguishing emotions with more ambiguous facial cues. Overall, the model demonstrates high reliability and robustness, particularly in distraction-related categories critical to driver monitoring systems.

Table 1: Classification Report of the Se-ResNet50 Model

Class	Precision	Recall	F1-Score	Support
Anger	0.77	0.75	0.76	500
Cognitively Distracted	0.99	1.00	1.00	500
Drowsy	1.00	1.00	1.00	500
Fear	0.87	0.81	0.84	500
Fidgetiness	1.00	1.00	1.00	1238
Neutral	0.74	0.80	0.77	500
Visually Distracted	1.00	1.00	1.00	500
Accuracy	—	—	0.92	4238
Macro Avg	0.91	0.91	0.91	4238
Weighted Avg	0.92	0.92	0.92	4238

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

Table 2: Classification Report of the Hybrid CNN-SE-BiGRU Attention Network

Class	Precision	Recall	F1-Score	Support
Anger	0.98	0.98	0.98	385
Fear	1.00	0.98	0.99	352
Neutral	0.94	0.97	0.95	261
Accuracy	—	—	0.98	998
Macro Avg	0.97	0.98	0.97	998
Weighted Avg	0.98	0.98	0.98	998

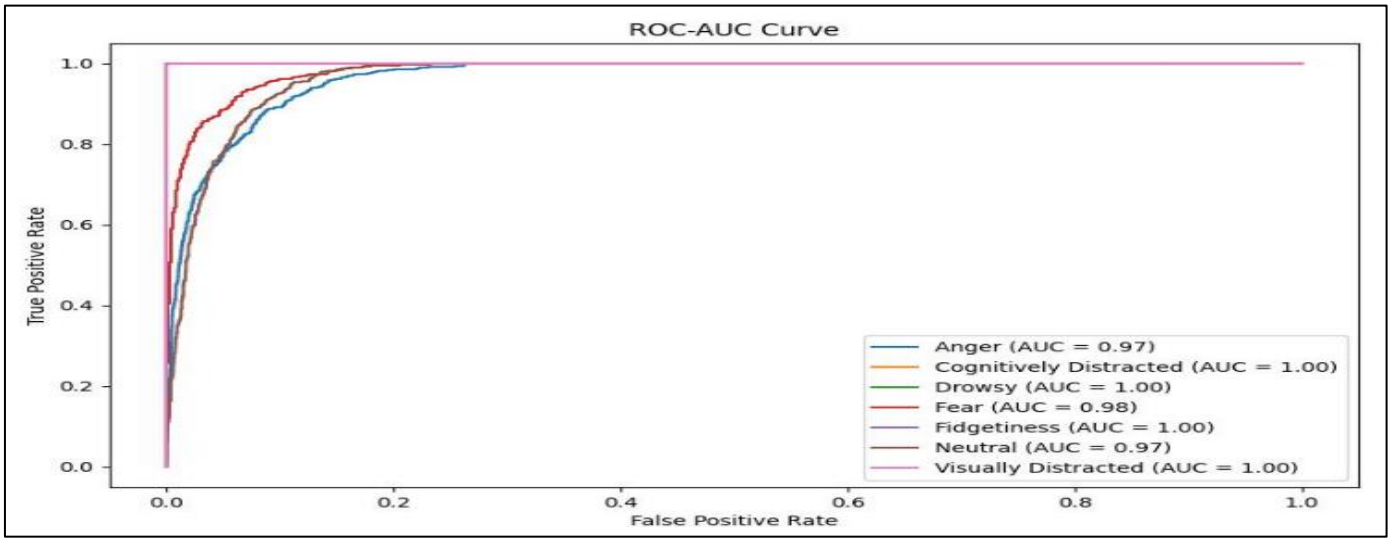

Fig 6 ROC-AUC Curves for Se-ResNet50 Model

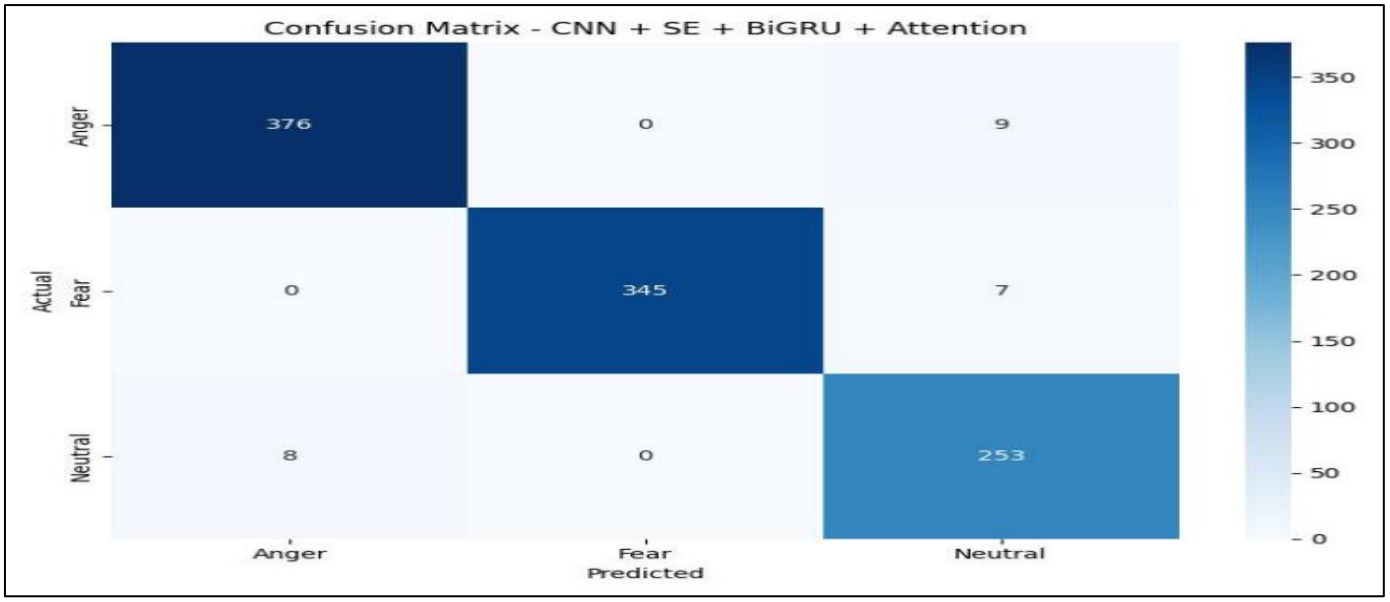

Fig 7 Confusion Matrix of the Hybrid CNN-SE-BiGRU Attention Network

The results of the CNN-BiGRU model are presented in Table 2. The *CNN-BiGRU model*, enhanced with an *attention mechanism*, exhibits superior performance metrics. It achieves a *weighted average F1-score* of 0.9761 and an overall accuracy of 98%, indicating a high level of predictive capability. The *ROC-AUC score* stands at an impressive 0.9987, reflecting excellent class discrimination. Among individual classes, *Fear* stands out as the best-performing, achieving perfect precision (no false positives), a recall of 98% (only 7 out of 352 samples misclassified), and an F1-score of 0.99. *Anger* also performs strongly, with both precision and recall around 0.98, though it shows some overlap with the *Neutral* class, as evidenced by 9 instances being misclassified as *Neutral*.

The *Neutral* class maintains high recall at 0.97, ensuring most true cases are correctly identified, but has a slightly lower precision of 0.94, suggesting a higher rate of false positives—mostly due to confusion with the *Anger* class.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

The *confusion matrix* for the *CNN-BiGRU model* reveals a notable interaction between the *Anger* and *Neutral* classes (Fig. 7). Specifically, 9 *Anger* samples are misclassified as *Neutral*, and 8 *Neutral* samples are misclassified as *Anger*. This bidirectional misclassification suggests a degree of feature similarity or overlaps between *neutral* and *angry* speech patterns. In contrast, the *Fear* class appears to be well-separated from the others, with minimal confusion—only 7 instances are misclassified, all as *Neutral*—highlighting its distinctiveness in the feature space.

These results underscore the effectiveness of incorporating *temporal dynamics* and *attention mechanisms* in modeling driver behaviors. The *CNN-BiGRU model*'s superior performance suggests that capturing sequential dependencies and focusing on relevant features are crucial for accurately identifying complex emotional and behavioral states. The high accuracy and F1-scores across most classes indicate that the model is well-suited for *real-time driver monitoring* applications, where timely and accurate detection of driver states is essential for ensuring *road safety*.

During testing, each second of the input video is processed in real time. From each second, a representative frame is extracted and passed through the Se-ResNet50 model for visual emotion and behavioral state classification. Simultaneously, the corresponding audio segment of that second is sampled. For the first second, the one-second audio segment is directly processed using the CNN-SE-BiGRU model. From the second-second onward, a sliding window approach with a window size of two seconds is applied. For example, at the second second, the model considers the audio from the first and second seconds; at the third second, it uses audio from the second and third seconds, and so on. This strategy enhances temporal context and emotional consistency in speech analysis. All operations are performed in real time during testing.

In addition to video and audio inputs, the system also incorporates vehicle speed telemetry data corresponding to each video frame. A priority-based alerting mechanism is defined to handle conflicting or uncertain predictions between modalities. Each predefined emotion or behavioral action is associated with a threshold probability for raising a warning. If the same emotion or behavior is detected by both the video and audio models, the model with the higher probability score determines the modality for issuing a voice warning. If the detected classes differ between the two models, the decision defaults to the video-based (Se-ResNet50) model output.

In case of a tie in probability scores across multiple classes, a predefined priority hierarchy governs the final warning decision. The priority levels for warnings, based on safety relevance and supported modality, are outlined as Table 3. The system logs probability distributions for all classes every second. An example of this output is shown below, providing a visual overview of how driver behaviors evolve over time.

Table 3: Priority Levels for Warnings Based on Behavior/Emotion and Input Modality

Priority	Behavior/Emotion	Modality
1	Speed of the Vehicle	Vehicle Data
2	Drowsy	Image Only
3	Visually Distracted	Image Only
4	Cognitively Distracted	Image Only
5	Anger	Image and Audio
6	Fidgetiness	Image Only
7	Fear	Image and Audio

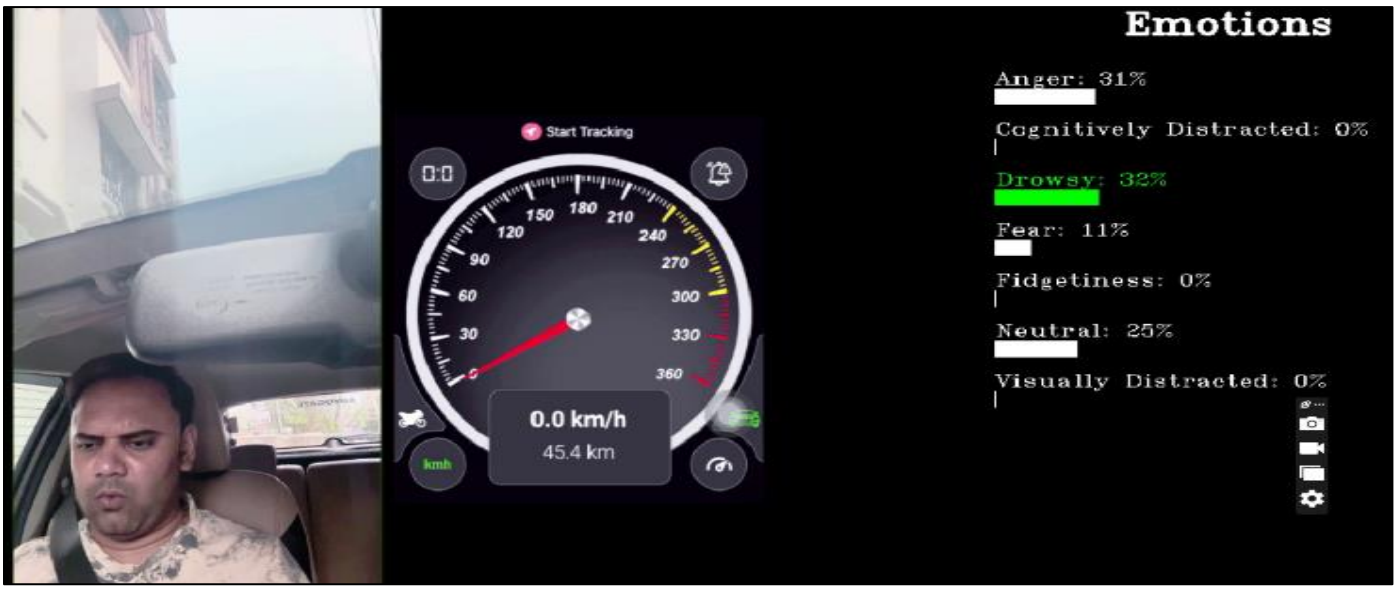

Fig 8 Real-Time Probability Distribution of Predicted Driver Behaviors and Emotions over Each Second of the Test Video

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

Table 4 Comparison of Video-Based and Audio-Based Emotion Recognition Models

Domain	Paper (Authors, Year)	Dataset Used	Algorithm Used	Results
Video	Huang *et al.* [3]	AffectNet RAF-DB	SE-ResNet CNN	Accuracy: 56.54% Accuracy: 83.37%
	Proposed Work	AffectNet Drowsy Detection State Farm	Se-ResNet50	Accuracy: 92% F1-score: 0.92
Audio	Liu *et al.* [4]	IEMOCAP	CNN-GRU-Attention	UA: 70.42%, WA: 67.79%
	Proposed Work	RAVDESS CREMA-D SAVEE, TESS	CNN-SE-BiGRU with Attention	Accuracy: 98% F1-score: 0.9761 ROC-AUC: 0.9987

The proposed system's visualization is shown in Fig. 8. This visualization serves as qualitative validation for the model's real-time capabilities and reveals transitions in emotional or behavioral states—such as from Neutral to Anger or Drowsiness—captured in the field test.

The performance of the current unimodal models provides a strong foundation for the next phase of development. The upcoming multimodal classifier will be evaluated using the same metrics, with a focus on reduced false positives in critical categories such as Drowsy and Visually Distracted. Comparative results will be included in future iterations of this work once the classifier is trained and validated.

VI. EVALUATION AGAINST STATE-OF-THE-ART APPROACHES

To evaluate the performance of the proposed models, a systematic comparison is carried out with methods reported in related literature. Table 4 presents a comparison of the video-based and audio-based emotion recognition models with state-of-the-art methods available in the literature. This comparison is designed to place our results in the context of existing research and to assess whether the proposed approach offers improvements over established techniques. The studies selected for comparison address similar tasks and employ comparable methodological frameworks. The evaluation considers several key aspects: the datasets used, the underlying model architectures, and the performance metrics reported. Dataset characteristics, including size, diversity, and class distribution, are carefully reviewed to ensure a fair basis for comparison. Similarly, architectural differences, such as the type and depth of neural network layers or the inclusion of attention mechanisms, are examined, as these factors can significantly affect outcomes. Performance is evaluated using commonly adopted metrics such as accuracy, and F1-score. These metrics provide complementary perspectives, helping to avoid bias toward any single aspect of performance.

In addition to accuracy-oriented measures, considerations such as computational efficiency, memory usage, and inference time were considered, as these can be critical for real-world applications. By analyzing these dimensions together, this evaluation aims to present a balanced and transparent assessment of how the proposed models compare to the state of the art. The insights gained from this process not only demonstrate the strengths of the proposed approach but also highlight areas for further improvement.

VII. CONCLUSION

This research proposes a *multimodal system* for predictive driver behavior monitoring and real-time intervention, integrating both video and audio inputs to detect emotional and behavioral states such as drowsiness, distraction, fear, and anger. The Se-ResNet50 image-based model achieved an accuracy of 92% with a weighted average F1-score of 0.92, showing strong performance in detecting states like *cognitively distracted*, *drowsy*, and *visually distracted*. The hybrid CNN-SE-BiGRU audio model attained an accuracy of 98%, a weighted F1-score of 0.9761, and a ROC-AUC of 0.9987, indicating excellent class discrimination for *anger*, *fear*, and *neutral* emotional states. The system also incorporates vehicle speed telemetry to provide contextual cues, and real-time tests validate its responsiveness and interpretability in dynamic driving scenarios.

REFERENCES

[1]. World Health Organization, "Global status report on road safety 2023," World Health Organization Publications, 2023.
[2]. F. Qu, N. Dang, B. Furht, and M. Nojoumian, "Comprehensive study of driver behavior monitoring systems using computer vision and machine learning techniques," Journal of Big Data, vol. 11, p. 32, 2024.
[3]. Z.-Y. Huang et al., "A study on computer vision for facial emotion recognition," Scientific Reports, 2023.
[4]. G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," EURASIP Journal on Audio, Speech, and Music Processing, p. 22, 2023.
[5]. R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial expression recognition in videos using hybrid CNN and ConvLSTM," International Journal of Information Technology, vol. 15, no. 4, pp. 1819–1830, 2023.
[6]. M. Mohana, P. Subashini, and M. Krishnaveni, "Emotion recognition from facial expressions using

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar176

hybrid CNN–LSTM network," International Journal of Pattern Recognition and Artificial Intelligence, vol. 37, no. 8, 2023.

[7]. B. Niu, Z. Gao, and B. Guo, "Facial expression recognition with LBP and ORB features," Computational Intelligence and Neuroscience, 2021.

[8]. Y. Albadawi, M. Takruri, and M. Awad, 'A review of recent developments in driver drowsiness detection systems', *Sensors*, vol. 22, no. 5, p. 2069, 2022.

[9]. J. Hu, L. Shen, and G. Sun, 'Squeeze-and-excitation networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10]. M. M. Kabir, T. A. Anik, M. S. Abid, M. F. Mridha, and M. A. Hamid, "Facial expression recognition using CNN-LSTM approach," in IEEE international conference on science & contemporary technologies (ICSCT), 2021.

[11]. M. Selvaraj, R. Bhuvana, and S. Padmaja, "Human speech emotion recognition," International Journal of Engineering and Technology, vol. 8, no. 1, 2016.

[12]. H. Aouani and Y. B. Ayed, "Speech emotion recognition with deep learning," Procedia Computer Science, vol. 176, pp. 251–260, 2020.

[13]. A. Kumar, "A new fitness function in genetic programming for classification of imbalanced data," Journal of Experimental & Theoretical Artificial Intelligence, vol. 36, no. 7, pp. 1021–1033, 2024.

[14]. A. Kumar, P. Maurya, S. M. Tiwari, A. Ali, H. Vasisht, and A. S. Baghel, "Classification of forest cover-type using ensemble of decision tree, random forest and k nearest neighbor," JIMS8I International Journal of Information Communication and Computing Technology, vol. 10, no. 2, pp. 615–619, 2022.

[15]. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE transactions on affective computing, vol. 10, no. 1, pp. 18–31, 2017.

[16]. A. Montoya, D. Holman, SF_data_science, T. Smith, and W. Kan, "State Farm Distracted Driver Detection,". https://kaggle.com/competitions/state-farm-distracted-driver-detection.